APPROXIMATION OF PROJECTIONS OF RANDOM VECTORS

ELIZABETH MECKES

Abstract. Let X be a d-dimensional random vector and X_θ its projection onto the span of a set of orthonormal vectors $\{\theta_1, \ldots, \theta_k\}$. Conditions on the distribution of X are given such that if θ is chosen according to Haar measure on the Stiefel manifold, the bounded-Lipschitz distance from X_θ to a Gaussian distribution is concentrated at its expectation; furthermore, an explicit bound is given for the expected distance, in terms of d, k, and the distribution of X, allowing consideration not just of fixed k but of k growing with d. The results are applied in the setting of projection pursuit, showing that most k-dimensional projections of n data points in $\mathbb{R}^d$ are close to Gaussian, when n and d are large and $k = c\sqrt{\log(d)}$ for a small constant c.

1. Introduction

There is a large class of results dealing with random variables (or measures) defined in terms of a parameter (say, a point on the sphere), which say that for a large measure of these parameters, the behavior of the random variable is well-approximated by some model distribution. Early work in this direction was done by Sudakov [21], who showed that under some relatively mild conditions, most one-dimensional marginals of a high-dimensional measure are close to each other. This line of research was further developed by von Weizsäcker [23], who showed that the canonical distribution around which one-dimensional marginals tend to cluster is close to a mixture of Gaussian distributions. In both [21] and [23], the results are about the limiting behavior of one-dimensional projections, as the ambient dimension tends to infinity, although von Weizsäcker points out that one could extend the methods to deal with higher fixed-dimensional projections, as the ambient dimension tends to infinity. More recent work in this area was done by Bobkov [3], who obtained concentration results for the distance from a one-dimensional projection of an isotropic log-concave random vector to a Gaussian distribution.

The purpose of this paper is to prove multivariate versions of such theorems; that is, to consider rank k projections of random vectors, instead of just rank one. Moreover, the approach yields results of a sufficiently quantitative nature to allow not only k fixed, but k growing with the ambient dimension. The general case of approximating random k-dimensional projections of probability measures on $\mathbb{R}^d$ is considered, and is illustrated with an application to graphical projection pursuit. In particular, it is shown that typical k-dimensional projections of n data points in $\mathbb{R}^d$ are close to Gaussian for n and d large; the precise quantitative nature of the results yields limit theorems even for $k = c\log(d)$ for a small constant c. This result generalizes the following univariate limit result of Diaconis and Freedman.

Theorem 1 (Diaconis-Freedman [7]). *Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$. Suppose that n, d and the x_i depend on a hidden index ν, so that as ν tends to infinity, so do n and*

Research supported by an American Institute of Mathematics Five-year Fellowship and NSF grant DMS-0852898.

d. Suppose that there is a $\sigma^2 > 0$ *such that, for all* $\epsilon > 0$,

$$\frac{1}{n}\Big|\left\{j \le n : \big||x_j|^2 - \sigma^2 d\big| > \epsilon d\right\}\Big| \xrightarrow{\nu\to\infty} 0, \tag{1}$$

and suppose that

$$\frac{1}{n^2}\Big|\left\{j, k \le n : \big|\langle x_j, x_k\rangle\big| > \epsilon d\right\}\Big| \xrightarrow{\nu\to\infty} 0. \tag{2}$$

Let $\theta \in \mathbb{S}^{d-1}$ *be distributed uniformly on the sphere, and consider the random measure* μ_ν^θ *which puts mass* $\frac{1}{n}$ *at each of the points* $\langle\theta, x_1\rangle, \ldots, \langle\theta, x_n\rangle$. *Then as* ν *tends to infinity, the measures* μ_ν^θ *tend to* $\mathfrak{N}(0, \sigma^2)$ *weakly in probability.*

The method of proof here is described in a fairly specific context: random measures indexed by points in the Stiefel manifold (one could equivalently take points in the Grassman manifold), approximated by Gaussian distributions. However, the approach is quite general and could in principle be adapted to a family of random measures indexed by points in a metric probability space possessing the concentration of measure phenomenon. Further, one could easily adapt the program to deal with non-Gaussian limits. In particular, Stein's method has been used to prove approximation results for many other limiting distributions, e.g. Poisson [5, 1, 2]; gamma [14]; chi-square [18]; uniform on the discrete circle [6]; the semi-circle law [10]; the binomial and multinomial distributions [11, 13]; and the hypergeometric distribution [11]; these approaches could be combined with what is done here in order to approximate by non-Gaussian distributions.

Before outlining the approach, some notation is needed. The Euclidean length of a vector $x \in \mathbb{R}^k$ is denoted $|x|$. For an $n \times n$ matrix $M = \big[m_{ij}\big]_{i,j=1}^n$, the Hilbert-Schmidt norm is defined by

$$\|M\|_{HS} = \mathrm{Tr}\,(MM^T) = \sqrt{\sum_{i,j} m_{ij}^2}.$$

The Wasserstein distance between two random vectors X and Y is defined by

$$d_W(X, Y) = \sup_{\{f : |f(x)-f(y)| \le |x-y|\}} \big|\mathbb{E}f(Y) - \mathbb{E}f(X)\big|.$$

The bounded-Lipschitz distance is defined by

$$d_{BL}(X, Y) := \sup_{\|f\|_1 \le 1} \big|\mathbb{E}f(S) - \mathbb{E}f(Y)\big|,$$

where

$$\|f\|_1 := \max\left\{\|f\|_\infty, \sup_{x \ne y} \frac{|f(x) - f(y)|}{|x - y|}\right\}.$$

The class of m-times continuously differentiable functions on $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted $C^m(\mathcal{X})$, and has a norm defined by

$$\|f\|_m := \sup_{0 \le k \le m} \sup_{x \in \mathcal{X}} \|D^k f(x)\|_{op}.$$

Here, $D^k f(x)$ denotes the symmetric k-linear form given in components by

$$D^k f(x)(y_1, \ldots, y_k) := \sum_{i_1, \ldots, i_k} \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x) y_1^{i_1} \cdots y_k^{i_k},$$

where $y_j = (y_j^1, \ldots, y_j^d)$. For an intrinsic definition of $D^k f(x)$, see Federer [9]. The ball of radius R in $C^m(\mathfrak{X})$ with respect to $\|\cdot\|_m$ is denoted $C_R^m(\mathfrak{X})$.

The Stiefel manifold $\mathfrak{W}_{d,k}$ is defined by

$$\mathfrak{W}_{d,k} = \{\theta = (\theta_1, \ldots, \theta_k) : \theta_i \in \mathbb{R}^d, \langle \theta_i, \theta_j \rangle = \delta_{ij} \,\forall\, 1 \leq i, j \leq k\},$$

with metric $\rho\big(\theta, \theta'\big) = \left[\sum_{j=1}^k |\theta_j - \theta_j'|^2\right]^{1/2}$. There is a unique rotation-invariant probability measure (Haar measure) on $\mathfrak{W}_{d,k}$; one way to construct it is by choosing θ_1 uniformly from $\mathbb{S}^{d-1}$, then θ_2 uniformly from the orthogonal complement of θ_1 in $\mathbb{S}^{d-1}$, and so on.

Now, suppose that a family of random vectors X_θ in $\mathbb{R}^k$ is indexed by $\theta \in \mathfrak{W}_{d,k}$. The following is an outline of an approach to show that most X_θ are approximately Gaussian.

1. Prove an approximation result for the average distribution. If X_θ is defined fairly explicitly in terms of θ, one can first try to use the following abstract normal approximation theorem to show that the average distribution of the X_θ (averaged over θ distributed according to Haar measure on $\mathfrak{W}_{d,k}$) is close to Gaussian.

Theorem 2 ([4]). *Let X be a random vector in $\mathbb{R}^k$ and for each $\epsilon > 0$ let X_ϵ be a random vector such that $\mathcal{L}(X) = \mathcal{L}(X_\epsilon)$, with the property that $\lim_{\epsilon \to 0} X_\epsilon = X$ almost surely. Let Z be a standard normal random vector in $\mathbb{R}^k$. Suppose there is a function $\lambda(\epsilon)$ and a random matrix F such that the following conditions hold.*

(i)

$$\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[(X_\epsilon - X)_i \middle| X\right] \xrightarrow[\epsilon \to 0]{L_1} -X.$$

(ii)

$$\frac{1}{2\lambda(\epsilon)}\mathbb{E}\left[(X_\epsilon - X)(X_\epsilon - X)^T | X\right] \xrightarrow[\epsilon \to 0]{L_1} \sigma^2 I_k + \mathbb{E}\left[F \middle| X\right].$$

(iii) *For each $\rho > 0$,*

$$\lim_{\epsilon \to 0} \frac{1}{\lambda(\epsilon)}\mathbb{E}\left[\big|X_\epsilon - X\big|^2 \mathbb{1}(|X_\epsilon - X|^2 > \rho)\right] = 0.$$

Then

$$d_W(X, \sigma Z) \leq \frac{1}{\sigma}\mathbb{E}\|F\|_{H.S.} \tag{3}$$

It should be pointed out that while this theorem is sufficiently general for the applications carried out here, there is a more general version (see [19] or [15]) allowing for approximations by Gaussian distributions with non-trivial covariance matrices. Furthermore, condition (i) need only hold approximately; see [4].

In order to apply this theorem, an auxiliary random variable $X_{\theta,\epsilon}$ must be constructed. A natural construction which makes use of the symmetry of $\mathfrak{W}_{d,k}$ is to let θ_ϵ be a "small random rotation" of θ (this is made explicit in the applications to follow). Then $(\theta, \theta_\epsilon)$ is an exchangeable pair of random points of $\mathfrak{W}_{d,k}$ by the rotation invariance of the distribution of θ, and so the random variables $(X_\theta, X_{\theta_\epsilon})$ are also exchangeable and thus have the same distribution. Furthermore, as $\epsilon \to 0$, $\theta_\epsilon \to \theta$ almost surely, and so if X_θ is a continuous function of θ, it will be true that $X_{\theta_\epsilon} \xrightarrow{\epsilon \to 0} X_\theta$ almost surely.

2. Use the concentration of measure on $\mathfrak{W}_{d,k}$ to show that for some distance $d(\cdot,\cdot)$, $d(X_\theta,\sigma Z)$ is close to its mean. It is shown in [17] that for $c_1 = \sqrt{\frac{\pi}{2}}$ and $c_2 = \frac{1}{8}$, for any $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ with median M_F and modulus of continuity $\omega_F(\eta)$,

$$\mathbb{P}\big[|F(\theta_1,\dots,\theta_k) - M_F| > \omega_F(\eta)\big] < c_1 e^{-c_2\eta^2 d}. \tag{4}$$

Here, $\mathbb{P}$ is the rotation-invariant probability measure on $\theta_1,\dots,\theta_k$ described above. The median M_F is a median with respect to this measure.

Again, if the random variable X_θ is a sufficiently regular function of θ, this theorem can be applied to the function $F(\theta) = d_{BL}(X_\theta,\sigma Z)$, where $d_{BL}(X_\theta,\sigma Z)$ is the conditional bounded-Lipschitz distance from X_θ to σZ, given θ. Standard arguments allow the median M_F to be replaced by the mean $\mathbb{E}F(\theta)$, with only minor loss.

3. Use entropy methods to bound $\mathbb{E}d_{BL}(X_\theta,\sigma Z)$. Consider the stochastic process $Y_f := \big|\mathbb{E}_X f(X_\theta) - \mathbb{E}f(X_\theta)\big|$ indexed by the class of functions $\{f : \|f\|_1 \le 1\}$ (or by some sub-class), where $\mathbb{E}_X$ denotes expectation with respect to X only; that is, conditional expectation with respect to the distribution of X, θ, conditioned on θ. Thus the bounded-Lipschitz distance from X_θ (given θ) to its average distribution can be viewed as the supremum of a stochastic process. The same approach used to prove a concentration result for $d_{BL}(X_\theta,\sigma Z)$ can be used to show that Y_f satisfies a sub-Gaussian increment condition of the type

$$\mathbb{P}\left[|Y_f - Y_g| > \epsilon\right] \le c_1 e^{-\frac{c_2\epsilon^2}{\|f-g\|_1^2}},$$

for some constants c_1 and c_2. For such a process, Dudley's entropy bound can be used to estimate its supremum. Specifically, Dudley showed the following.

Theorem 3 (Dudley, [8])**.** *Let $\{X_t\}_{t\in T}$ be a stochastic process indexed by a metric space T with distance d. Suppose that there is a constant c such that X_t satisfies the increment condition*

$$\forall u, \quad \mathbb{P}\left[|X_t - X_s| \ge u\right] \le c\exp\left(-\frac{u^2}{2d(s,t)^2}\right).$$

Then there is a constant C such that

$$\mathbb{E}\sup_{t\in T} X_t \le C\int_0^\infty \sqrt{\log N(T,d,\epsilon)}\,d\epsilon,$$

where $N(T,d,\epsilon)$ is the ϵ-covering number of T with respect to the distance d.

One can apply this theorem not to the index set $\{f : \|f\|_1 \le 1\}$ (which has infinite ϵ-covering number with respect to $\|\cdot\|_1$ for $\epsilon < 2$), but to a more restricted indexing set $\mathcal{F}$ of test functions. One may then be able to obtain a bound on $\mathbb{E}d_{BL}(X_\theta,\sigma Z)$ by approximation of functions f with $\|f\|_1 \le 1$ by functions from $\mathcal{F}$, together with the approximation for the average distribution proved in Step 1.

2. Random Projections

In this section, the method outlined in the introduction is applied in the case that X is a random vector in $\mathbb{R}^d$, $\theta = (\theta_1,\dots,\theta_k) \in \mathfrak{W}_{d,k}$, and X_θ is the projection of X onto the span of θ; that is

$$X_\theta := \big(\langle X,\theta_1\rangle,\dots,\langle X,\theta_k\rangle\big).$$

If θ is chosen randomly from $\mathfrak{W}_{d,k}$ (according to the rotation-invariant probability measure described in the introduction), then the distributions of the X_θ are a family of random measures on $\mathbb{R}^k$ indexed by θ.

To apply the method of the introduction, consider the random variable X_θ defined above, in the case that θ is chosen at random and independent of X. The following results describe the behavior of X_θ, both on average and conditioned on θ.

Theorem 4. *Let X be a random vector in $\mathbb{R}^n$, with $\mathbb{E}X = 0$, $\mathbb{E}\left[|X|^2\right] = \sigma^2 d$, and $\mathbb{E}\big||X|^2\sigma^{-2} - d\big| := A < \infty$. If θ is a random point of $\mathfrak{W}_{d,k}$ and X_θ is defined above,*

$$d_W(X_\theta, \sigma Z) \leq \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.$$

Theorem 5. *Suppose that B is defined by $B := \sup_{\xi \in \mathbb{S}^{d-1}} \mathbb{E}\left\langle X, \xi \right\rangle^2$. For $\theta \in \mathfrak{W}_{d,k}$, let*

$$d_{BL}(X_\theta, \sigma Z) = \sup_{\|f\|_1 \leq 1} \left|\mathbb{E}\left[f(\langle X, \theta_1\rangle, \ldots, \langle X, \theta_k\rangle)\middle|\theta\right] - \mathbb{E}f(\sigma Z_1, \ldots, \sigma Z_k)\right|;$$

that is, $d_{BL}(X_\theta, \sigma Z)$ is the conditional bounded-Lipschitz distance from X_θ to σZ, conditioned on θ. Then for $\epsilon > 2\pi\sqrt{\frac{B}{d}}$, and θ a random point of $\mathfrak{W}_{d,k}$

$$\mathbb{P}\left[\left|d_{BL}(X_\theta, \sigma Z) - \mathbb{E}d_{BL}(X_\theta, \sigma Z)\right| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}e^{-\frac{d\epsilon^2}{32B}}.$$

Theorem 6. *There is a constant $C > 1$ such that*

$$\mathbb{E}d_{BL}(X_\theta, \sigma Z) \leq \frac{C^k B}{d^{\frac{2}{9k+4}}} + \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.$$

Observe that together, Theorems 5 and 6 show that for $\epsilon \geq \frac{2C^k B}{d^{\frac{2}{9k+4}}} + \frac{2\sigma\sqrt{k}(A+1) + 2\sigma k}{d-1}$,

$$\mathbb{P}\left[d_{BL}(X_\theta, \sigma Z) > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}e^{-\frac{d\epsilon^2}{2^7 B}}.$$

Note that the bound on the right tends to zero as $d \to \infty$ for any ϵ in this range.

Proof of Theorem 4. Observe first that $\mathbb{E}X_\theta = 0$ by symmetry and

$$\mathbb{E}(X_\theta)_i(X_\theta)_j = \mathbb{E}\left\langle \theta_i, X\right\rangle\left\langle \theta_j, X\right\rangle = \sum_{r,s=1}^{d} \mathbb{E}\left[\theta_{ir}\theta_{js}\right]\mathbb{E}\left[X_r X_s\right] = \frac{\delta_{ij}}{d}\mathbb{E}\left[|X|^2\right] = \delta_{ij}\sigma^2,$$

where the second-last equality follows from $\mathbb{E}\left[\theta_{ir}\theta_{js}\right] = \frac{1}{d}\delta_{ij}\delta_{rs}$.

To apply the Theorem 2 to X_θ, one first has to construct $X_{\theta,\epsilon}$. Let

$$A_\epsilon := \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus I_{d-2} = I_d + \begin{bmatrix} -\frac{\epsilon^2}{2} + \delta & \epsilon \\ -\epsilon & -\frac{\epsilon^2}{2} + \delta \end{bmatrix} \oplus 0_{d-2},$$

where $\delta = O(\epsilon^4)$. Let $U \in \mathcal{O}_d$ be a random orthogonal matrix, independent of X, and define $X_{\theta,\epsilon} := \left(\left\langle UA_\epsilon U^T\theta_1, X\right\rangle, \ldots, \left\langle UA_\epsilon U^T\theta_k, X\right\rangle\right)$; the pair $(X_\theta, X_{\theta,\epsilon})$ is exchangeable by the rotation invariance of the distribution of θ, and so $\mathcal{L}(X_\theta) = \mathcal{L}(X_{\theta,\epsilon})$.

Let K be the $d \times 2$ matrix given by the first two columns of U and let $C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$; define the matrix $Q = \big[q_{ij}\big]_{i,j=1}^{d} = KCK^T$. Then, writing $X_\theta = (X_1^\theta, \ldots, X_k^\theta)$ and $X_{\theta,\epsilon} = (X_{\epsilon,1}^\theta, \ldots, X_{\epsilon,k}^\theta)$,

$$\begin{aligned}\mathbb{E}\left[X_{\epsilon,j}^\theta - X_j^\theta \middle| X, \theta\right] &= \mathbb{E}\left[\left\langle (UA_\epsilon U^T - I)\theta_j, X\right\rangle \middle| X, \theta\right] \\ &= \epsilon \mathbb{E}\left[\langle Q\theta_j, X\rangle \middle| X, \theta\right] - \frac{\epsilon^2}{2}\mathbb{E}\left[\left\langle KK^T\theta_j, X\right\rangle \middle| X, \theta\right] + O(\epsilon^4).\end{aligned}$$

Recall that Q and K are determined by U alone, and that U is independent of X, θ. It is easy to show that $\mathbb{E}\big[Q\big] = 0_d$ and $\mathbb{E}\big[KK^T\big] = \frac{2}{d}I_d$, thus

$$\mathbb{E}\left[X_{\theta,\epsilon} - X_\theta \middle| X, \theta\right] = -\frac{\epsilon^2}{d}X_\theta + O(\epsilon^4).$$

Condition (i) of Theorem 2 is thus satisfied with $\mathcal{F} = \sigma(X, \theta)$ and $\lambda(\epsilon) = \frac{\epsilon^2}{d}$.

It is elementary but tedious to show that $\mathbb{E}q_{rs}q_{tv} = \frac{2}{d(d-1)}\big[\delta_{rt}\delta_{sv} - \delta_{rv}\delta_{st}\big]$ (the computation is carried out in detail in [4]). Making use of this yields

$$\begin{aligned}&\mathbb{E}\left[(X_{\epsilon,j}^\theta - X_j^\theta)(X_{\epsilon,\ell}^\theta - X_\ell^\theta)\middle| X, \theta\right] \\ &\qquad = \epsilon^2 \mathbb{E}\left[\langle Q\theta_j, X\rangle \langle Q\theta_\ell, X\rangle \middle| X, \theta\right] + O(\epsilon^3) \\ &\qquad = \epsilon^2 \sum_{r,s,t,v=1}^{d} \mathbb{E}\left[q_{rs}q_{tv}\theta_{js}\theta_{\ell v}X_rX_t \middle| X, \Theta\right] + O(\epsilon^3) \\ &\qquad = \frac{2\epsilon^2}{d(d-1)}\left[\sum_{r,s=1}^{d}\theta_{js}\theta_{\ell s}X_r^2 - \sum_{r,s=1}^{d}\theta_{js}\theta_{\ell r}X_rX_s\right] + O(\epsilon^3) \\ &\qquad = \frac{2\epsilon^2}{d(d-1)}\left[\delta_{j\ell}|X|^2 - X_j^\theta X_\ell^\theta\right] + O(\epsilon^3) \\ &\qquad = \frac{2\epsilon^2\sigma^2}{d}\delta_{j\ell} + \frac{2\epsilon^2}{d(d-1)}\left[\delta_{j\ell}\big(|X|^2 - \sigma^2 d\big) + \delta_{j\ell}\sigma^2 - X_j^\theta X_\ell^\theta\right] + O(\epsilon^3).\end{aligned}$$

The random matrix F of Theorem 2 is thus defined by

$$F = \frac{1}{d-1}\left[\big(|X|^2 - \sigma^2 d\big)I_k + \sigma^2 I_k - X_\theta X_\theta^T\right].$$

It follows from the theorem that

$$\begin{aligned}d_W(W, \sigma Z) &\le \frac{1}{\sigma}\mathbb{E}\|F\|_{H.S.} \\ &\le \frac{\sigma\sqrt{k}}{d-1}\left[\mathbb{E}\left|\frac{|X|^2}{\sigma^2} - d\right| + 1\right] + \frac{\sigma}{d-1}\mathbb{E}\left[\sum_j\left(\frac{X_j^\theta}{\sigma}\right)^2\right] \qquad (5) \\ &\le \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.\end{aligned}$$

□

Proof of Theorem 5. Define a function $F : \mathfrak{W}_{d,k} \to \mathbb{R}$ by

$$F(\theta) = \sup_{\|f\|_1 \le 1} \left| \mathbb{E}_X f(X_\theta)) - \mathbb{E} f(\sigma Z) \right|,$$

where $\mathbb{E}_X$ denotes the expectation with respect to the distribution of X only; that is,

$$\mathbb{E}_X f(X_\theta) = \mathbb{E}\left[f(X_\theta) \middle| \theta \right].$$

To apply the concentration of measure on $\mathfrak{W}_{d,k}$, it is necessary to determine the modulus of continuity of F. First, observe that for f with $\|f\|_1 \le 1$ given,

$$\begin{aligned} \Big| \left| \mathbb{E}_X f(X_\theta) - \mathbb{E} f(\sigma Z) \right| \; &- \; \left| \mathbb{E}_X f(X'_\theta) - \mathbb{E} f(\sigma Z) \right| \Big| \\ &\le \left| \mathbb{E}_X f(X'_\theta) - \mathbb{E}_X f(X_\theta)) \right| \\ &= \mathbb{E}\left[f\big(\langle X, \theta'_1 \rangle, \dots, \langle X, \theta'_k \rangle \big) - f\big(\langle X, \theta_1 \rangle, \dots, \langle X, \theta_k \rangle \big) \middle| \theta, \theta' \right] \\ &\le \mathbb{E}\left[\left| \big(\langle X, \theta'_1 - \theta_1 \rangle, \dots, \langle X, \theta'_k - \theta_k \rangle \big) \right| \middle| \theta, \theta' \right] \\ &\le \sqrt{\sum_{j=1}^k |\theta'_j - \theta_j|^2 \mathbb{E} \left\langle X, \frac{\theta'_j - \theta_j}{|\theta'_j - \theta_j|} \right\rangle^2} \\ &\le \rho(\theta, \theta') \sqrt{B}. \end{aligned}$$

It follows that

$$\begin{aligned} &\left| d_{BL}(X_\theta, \sigma Z) - d_{BL}(X_{\theta'}, \sigma Z) \right| \\ &\qquad = \left| \sup_{\|f\|_1 \le 1} \left| \mathbb{E}_X f(X_\theta) - \mathbb{E} f(\sigma Z) \right| - \sup_{\|f\|_1 \le 1} \left| \mathbb{E}_X f(X_{\theta'}) - \mathbb{E} f(\sigma Z) \right| \right| \\ &\qquad \le \sup_{\|f\|_1 \le 1} \Big| \left| \mathbb{E}_X f(X_\theta) - \mathbb{E} f(\sigma Z) \right| - \left| \mathbb{E}_X f(X_{\theta'}) - \mathbb{E} f(\sigma Z) \right| \Big| \\ &\qquad \le \rho(\theta, \theta') \sqrt{B}, \end{aligned}$$

thus $d_{BL}(X_\theta, \sigma Z)$ is a Lipschitz function on $\mathfrak{W}_{k,d}$, with Lipschitz constant $\sqrt{B}$. Applying the concentration of measure inequality from Inequality (4) of the introduction then implies that

$$\mathbb{P}\big[|F(\theta_1, \dots, \theta_k) - M_F| > \epsilon \big] < \sqrt{\frac{\pi}{2}} e^{-\frac{\epsilon^2 d}{8B}}.$$

Now, if $\theta = (\theta_1, \dots, \theta_k)$ is a Haar-distributed random point of $\mathfrak{W}_{d,k}$, then

$$\begin{aligned} \left| \mathbb{E} F(\theta) - M_F \right| \le \mathbb{E}\left| F(\theta) - M_F \right| = \int_0^\infty \mathbb{P}\Big[\left| F(\theta) - M_F \right| > t \Big] dt \\ \le \int_0^\infty \sqrt{\frac{\pi}{2}} e^{-\frac{dt^2}{8B}} dt = \pi \sqrt{\frac{B}{d}}. \end{aligned}$$

So as long as $\epsilon > 2\pi\sqrt{\frac{B}{d}}$, replacing the median of F with its mean only changes the constants:

$$\begin{aligned}\mathbb{P}\left[|F(\theta) - \mathbb{E}F(\theta)| > \epsilon\right] &\leq \mathbb{P}\left[\left|F(\theta) - M_F\right| > \epsilon - \left|M_F - \mathbb{E}F(\theta)\right|\right] \\ &\leq \mathbb{P}\left[|F(\theta) - M_F| > \frac{\epsilon}{2}\right] \leq \sqrt{\frac{\pi}{2}}e^{-\frac{d\epsilon^2}{32B}}.\end{aligned}$$

□

What has just been shown is that $d_{BL}(X_\theta, \sigma Z)$ is concentrated about its mean; it remains to give a bound for this mean (Theorem 6).

Proof of Theorem 6. As indicated in the introduction, Theorem 6 is proved making use of Dudley's entropy bound for bounding the expected value of the supremum of a stochastic process. Let $X_f := \left|\mathbb{E}_X f(X_\theta) - \mathbb{E}f(X_\theta)\right|$. Then $\{X_f\}_f$ is a stochastic process (each X_f is a random variable depending on θ) indexed by a family of functions f. The same type of concentration argument used above can be used to show that this process is sub-Gaussian.

Let $f : \mathbb{R}^k \to \mathbb{R}$ be Lipschitz with Lipschitz constant L and consider the function $G = G_f$ defined on $\mathfrak{W}_{d,k}$ by

$$G(\theta_1, \ldots, \theta_k) = \mathbb{E}_X f(X_\theta) = \mathbb{E}\left[f(\langle\theta_1, X\rangle, \ldots, \langle\theta_k, X\rangle)\middle|\theta\right].$$

The same argument as above shows that G is Lipschitz on $\mathbb{S}^{d-1}$ with Lipschitz constant $L\sqrt{B}$. It thus follows from (4) that

$$\mathbb{P}\left[|G(\theta) - M_G| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}e^{-\frac{d\eta^2}{8L^2B}},$$

and if $\epsilon > 2\pi\sqrt{\frac{L^2B}{d}}$,

$$\mathbb{P}\left[|G(\theta) - \mathbb{E}G(\theta)| > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}e^{-\frac{d\epsilon^2}{32L^2B}}. \tag{6}$$

Observe that, for θ a Haar-distributed random point of $\mathfrak{W}_{d,k}$, $\mathbb{E}G(\theta) = \mathbb{E}f(X_\theta)$, and so (6) can be restated as

$$\mathbb{P}\left[X_f > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}\exp\left[-\frac{d\epsilon^2}{2^7L^2B}\right].$$

Note that

$$\begin{aligned}\left|X_f - X_g\right| &= \left|\left|\mathbb{E}_X f(X_\theta) - \mathbb{E}f(X_\theta)\right| - \left|\mathbb{E}_X g(X_\theta) - \mathbb{E}g(X_\theta)\right|\right| \\ &\leq \left|\mathbb{E}_X(f-g)(X_\theta) - \mathbb{E}(f-g)(X_\theta)\right| = X_{f-g},\end{aligned}$$

thus for $\epsilon > 4\pi L(f-g)\sqrt{\frac{B}{d}}$, for $L(f-g)$ the Lipschitz constant of $f-g$,

$$\mathbb{P}\left[\left|X_f - X_g\right| > \epsilon\right] \leq \mathbb{P}\left[X_{f-g} > \epsilon\right] \leq \sqrt{\frac{\pi}{2}}\exp\left[\frac{-d\epsilon^2}{2^7[L(f-g)]^2B}\right] \leq \sqrt{\frac{\pi}{2}}\exp\left[\frac{-d\epsilon^2}{2^7\|f-g\|_1^2B}\right].$$

The condition on ϵ may be removed by replacing the factor of $\sqrt{\frac{\pi}{2}}$ in the bound above by, e.g., $3\sqrt{\frac{\pi}{2}}$. The process $\{X_f\}$ therefore satisfies the sub-Gaussian increment condition for the distance $d^*(f,g) := \frac{8\sqrt{B}}{\sqrt{d}}\|f-g\|_1$.

Consider the class $C_1^m(B_R)$ of functions f which are supported on $B_R := \{x \in \mathbb{R}^k : |x| \le R\}$ such that $\|f\|_m := \sup_{0\le j\le m}\sup_{x\in B_R}\|D^j f(x)\|_{op} \le 1$. It is proved in the appendix that for $\epsilon < 2$ and $m \ge 2$, the ϵ-covering number for this set with respect to the norm $\|\cdot\|_1$ is bounded by

$$\exp\left[\left((3\log(5) - \frac{m}{m-1}\log(\epsilon) + \frac{c_1}{\epsilon^{\frac{k}{m-1}}}\right)e^{k+m-2}\right]$$

with

$$c_1 = \frac{2\pi^{k/2}(R+1)^k\big((m+4)\log(2)\big)(5)^{\frac{k}{m-1}}}{k\Gamma\left(\frac{k}{2}\right)}.$$

It follows that the ϵ-covering number with respect to the distance d^* is bounded by

$$\exp\left[\left(3\log(5) - \frac{m}{(m-1)}\log\left(\frac{\epsilon\sqrt{d}}{8\sqrt{B}}\right)\right.\right.$$
$$\left.\left.+\frac{2\log(2)(m+4)\left[\sqrt{\pi}(R+1)\right]^k\left[40\sqrt{B}\right]^{\frac{k}{m-1}}}{k\Gamma\left(\frac{k}{2}\right)\left[\epsilon\sqrt{d}\right]^{\frac{k}{m-1}}}\right)e^{k+m-2}\right].$$

Since functions $f \in C_1^m(B_R)$ have in particular $\|f\|_1 \le 1$, this class also satisfies the sub-Gaussian increment condition with respect to the metric d^*. Note that the diameter of $C_1^m(B_R)$ with respect to d^* is bounded above by $16\sqrt{\frac{B}{d}}$. It follows from Dudley's entropy bound that there is a constant C such that $\mathbb{E}\Big[\sup_{f\in C_1^m(B_R)} X_f\Big]$ is bounded above by

$$Ce^{\frac{k+m-2}{2}}\int_0^{16\sqrt{\frac{B}{d}}}\sqrt{3\log(5) - \frac{m}{(m-1)}\log\left(\frac{\epsilon\sqrt{d}}{8\sqrt{B}}\right) + \frac{2\log(2)(m+4)\left[\sqrt{\pi}(R+1)\right]^k\left[40\sqrt{B}\right]^{\frac{k}{m-1}}}{k\Gamma\left(\frac{k}{2}\right)\left[\epsilon\sqrt{d}\right]^{\frac{k}{m-1}}}}\,d\epsilon.$$

Making the substitution $s = \frac{\epsilon\sqrt{d}}{8\sqrt{B}}$ then gives an upper bound of

$$C'e^{\frac{k+m-2}{2}}\sqrt{\frac{B}{d}}\int_0^2\sqrt{3\log(5) - \frac{m}{m-1}\log(s) + \frac{2\log(2)(m+4)\left[\sqrt{\pi}(R+1)\right]^k(5)^{\frac{k}{m-1}}}{k\Gamma\left(\frac{k}{2}\right)s^{\frac{k}{m-1}}}}\,ds$$

for another constant C'. Looking at the first two summands and the third separately, as long as $m > \frac{k}{2}+1$, this implies that there is an absolute constant C such that $\mathbb{E}\Big[\sup_{f\in C_1^m(B_R)} X_f\Big]$ is bounded by

$$\sqrt{\frac{B}{d}}\left(\frac{C^{k+m}R^{k/2}m^{3/2}}{(2m-k-2)\sqrt{k\Gamma\left(\frac{k}{2}\right)}}\right),$$

or, as will be needed in what follows,

$$\mathbb{E}\left[\sup_{f\in C_M^m(B_R)} X_f\right] \le \sqrt{\frac{B}{d}}\left(\frac{MC^{k+m}R^{k/2}m^{3/2}}{(2m-k-2)\sqrt{k\Gamma\left(\frac{k}{2}\right)}}\right). \tag{7}$$

From this bound, one can obtain a bound on $\mathbb{E}d_{BL}(X_\theta,)$ as follows. Let

$$\varphi_R(x) = \begin{cases} 1 & |x| \le R, \\ R+1-|x| & R \le |x| \le R+1, \\ 0 & R+1 \le |x|; \end{cases}$$

that is, φ_R is a radially symmetric cut-off function with $\|\varphi_R\|_1 \le 1$, supported on B_{R+1} and with $\varphi_R \equiv 1$ on B_R. For $f \in C_1^1(\mathbb{R}^k)$, let $f_R := f \cdot \varphi_R$. Then

$$\|f_R\|_1 = \max\left\{\sup_x |f(x)\varphi_R(x)|, \sup_x |f(x)\cdot\nabla\varphi_R(x) + \varphi_R(x)\nabla f(x)|\right\} \le 2.$$

Since $|f(x) - f_R(x)| = 0$ if $x \in B_R$ and $|f(x) - f_R(x)| \le 1$ for all $x \in \mathbb{R}^k$,

$$\left|\mathbb{E}_X f(X_\theta) - \mathbb{E}_X f_R(X_\theta)\right| \le \mathbb{P}\left[|X_\theta| > R\middle|\theta\right] \le \frac{1}{R^2}\sum_{i=1}^k \mathbb{E}\left[\langle X, \theta_i\rangle^2\right] \le \frac{Bk}{R^2},$$

and the same holds if $\mathbb{E}_X$ is replaced by $\mathbb{E}$. It follows that

$$|X_f - X_{f_R}| \le \frac{2Bk}{R^2}. \tag{8}$$

Next, let $\psi : \mathbb{R} \to \mathbb{R}$ be a C^∞ bump function, such that $0 \le \psi(y) \le 1$ for all y, $\psi(y) = 1$ for $-1 \le y \le 1$, $\psi(y) = 0$ for $|y| > 2$, and such that

$$\left|\frac{d^j\psi}{dy^j}(y)\right| \le C^j j^{2j} \tag{9}$$

for all $j \in \mathbb{N}$ (the existence of such a function is guarranteed by Theorem 1.4.2 of [12]). For $x \in \mathbb{R}^k$, define

$$\psi_t(x) = \frac{C(k)}{t^k}\psi\left(\frac{|x|}{t}\right),$$

where $C(k)$ is a constant depending only on k, such that $\int_{\mathbb{R}^k}\psi_t = 1$. Observe that it follows from the bounds (9) that

$$\|D^j\psi_t(x)\|_{op} \le \frac{C(k)C^j j^{2j}}{t^{k+j}}\mathbb{1}(t \le |x| \le 2t). \tag{10}$$

For $g \in C_2^1(\mathbb{R}^k)$, let $g_t(x) := g * \psi_t(x)$. Let Y_t be a random vector in $\mathbb{R}^k$ with density ψ_t, independent of X, θ. Then one can write

$$\mathbb{E}_X g_t(X_\theta) = \mathbb{E}_X g(X_\theta + Y_t),$$

and the same with $\mathbb{E}$ in place of $\mathbb{E}_X$. Since $g \in C_2^1(\mathbb{R}^k)$, it follows that

$$\max\left(\left|\mathbb{E}_X g(X_\theta) - \mathbb{E}_X g_t(X_\theta)\right|, \left|\mathbb{E}g(X_\theta) - \mathbb{E}g_t(X_\theta)\right|\right) \le 2\mathbb{E}|Y_t| \le 4t,$$

from which it follows that

$$|X_g - X_{g_t}| \le 8t. \tag{11}$$

Furthermore, by Young's inequality, for $j \leq m$,

$$\|D^j g_t(x)\|_{op} \leq \|g\|_\infty \int_{\mathbb{R}^k} \|D^j \psi_t(y)\|_{op} dy \leq \frac{2C(k)C^j j^{2j}}{t^{k+j}} \operatorname{vol}(B_{2t}) = \frac{2^{k+2}\pi^{k/2} C(k) C^j j^{2j}}{t^j k \Gamma\left(\frac{k}{2}\right)}.$$

Now, integrating in polar coordinates,

$$\frac{1}{C(k)} = \int_{\mathbb{R}^k} t^{-k} \psi\left(\frac{|x|}{t}\right) dx = \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)} \int_0^2 \psi(r) r^{k-1} dr \geq \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)} \int_0^1 r^{k-1} dr = \frac{2\pi^{k/2}}{k\Gamma\left(\frac{k}{2}\right)}.$$

It follows that

$$\|D^j g_t(x)\|_{op} \leq \frac{2^{k+1} C^j j^{2j}}{t^j}$$

for all $x \in \mathbb{R}^k$, and so $\|g_t\|_m \leq \frac{2^{k+1} C^m m^{2m}}{t^m}$. Finally, if g is supported on B_{R+1}, then it is easy to see that g_t is supported on B_{R+1+2t}.

It now follows from (7), (8) and (11) that

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in C_1^1(\mathbb{R}^k)} X_f\right] &\leq \mathbb{E}\left(\sup_{f \in C_1^1(\mathbb{R}^k)} \left[|X_f - X_{f_R}| + \left|X_{f_R} - X_{(f_R)_t}\right| + X_{(f_R)_t}\right]\right) \\
&\leq \frac{2Bk}{R^2} + 8t + \sqrt{\frac{B}{d}}\left(\frac{2^{k+1} m^{2m} C^{k+m} (R+1+2t)^{k/2} m^{3/2}}{(2m-k-2)t^m \sqrt{k\Gamma\left(\frac{k}{2}\right)}}\right) \qquad (12) \\
&\leq \frac{2Bk}{R^2} + 8t + \sqrt{\frac{B}{d}}\left(\frac{C^{k+m} m^{2m} R^{k/2} m^{3/2}}{(2m-k-2)t^m \sqrt{k\Gamma\left(\frac{k}{2}\right)}}\right).
\end{aligned}$$

Choosing $t = \frac{k}{R^2}$ yields

$$\mathbb{E}\left[\sup_{f \in C_1^1(\mathbb{R}^k)} X_f\right] \leq \frac{(2B+8)k}{R^2} + \sqrt{\frac{B}{d}}\left(\frac{C^{k+m} m^{2m} R^{2m+k/2} m^{3/2}}{(2m-k-2)k^m \sqrt{k\Gamma\left(\frac{k}{2}\right)}}\right). \qquad (13)$$

Now choosing $m = k$ and applying Stirling's formula to $\Gamma\left(\frac{k}{2}\right)$ yields

$$\mathbb{E}\left[\sup_{f \in C_1^1(\mathbb{R}^k)} X_f\right] \leq \frac{(2B+8)k}{R^2} + \sqrt{\frac{B}{d}}\left[\left(Ck^{3/4}\right)^k R^{9k/2}\right]. \qquad (14)$$

Setting $R = \left(\frac{d}{k^{\frac{3k}{2}-2}}\right)^{\frac{1}{9k+4}}$ yields

$$\mathbb{E}\left[\sup_{f \in C_1^1(\mathbb{R}^k)} X_f\right] \leq \frac{C^k B}{d^{\frac{2}{9k+4}}}. \qquad (15)$$

Finally, by Theorem 4 and (15),

$$\begin{aligned}
\mathbb{E} d_{BL}(X_\theta, \sigma Z) &\leq \mathbb{E}\left(\sup_{f \in C_1^1(\mathbb{R}^k)} \left[|\mathbb{E}_X f(X_\theta) - \mathbb{E} f(X_\theta)| + |\mathbb{E} f(X_\theta) - \mathbb{E} f(\sigma Z)|\right]\right) \\
&\leq \frac{C^k B}{d^{\frac{2}{9k+4}}} + \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.
\end{aligned}$$

□

3. Application: Projection Pursuit

In this section, the theorems of the previous section are applied to prove a quantitative, higher-dimensional version of a result of Diaconis and Freedman [7]. Let $x_1, \ldots, x_n$ be deterministic vectors in $\mathbb{R}^d$; write $x_i = (x_{i,1}, \ldots, x_{i,d})$. Define $\sigma > 0$ by the condition $\frac{1}{n}\sum_{i=1}^n |x_i|^2 = \sigma^2 d$, and define A and B by $A := \frac{1}{n}\sum_{i=1}^n \left|\sigma^{-2}|x_i|^2 - d\right|$ and $B := \sup_{\theta\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle\theta, x_i\rangle^2$. Observe that $\frac{\sigma^2 d}{n} \leq B \leq \sigma^2 d$. Also,if X is distributed uniformly over the points $\{x_i\}$, then these definitions of σ, A, and B correspond to those in the previous section.

Let $\theta = (\theta_1, \ldots, \theta_k)$ be a random point in $\mathfrak{W}_{d,k}$, distributed according to the rotation-invariant probability measure described in the introduction, and consider the family of random measures $\mu^\theta_{n,d,k}$ defined in terms of θ by

$$\mu^\theta_{n,d,k} := \frac{1}{n}\sum_{i=1}^n \delta_{(\langle\theta_1,x_i\rangle,\ldots,\langle\theta_k,x_i\rangle)}.$$

That is, $\mu^\theta_{n,d,k}$ puts equal mass at the projections of each of the x_i onto the span of $\theta_1, \ldots, \theta_k$.

In Diaconis and Freedman [7] it was shown that, in the case $k = 1$, the measures $\mu^\theta_{n,d,1}$ converge weakly in probability to Gaussian as n and d tend to infinity, under the conditions that, for some $\sigma^2 > 0$ such that, for all $\epsilon > 0$,

$$\frac{1}{n}\left|\left\{j \leq n : \left||x_j|^2 - \sigma^2 d\right| > \epsilon d\right\}\right| \xrightarrow{\nu\to\infty} 0, \tag{16}$$

and

$$\frac{1}{n^2}\left|\left\{j, k \leq n : \left|\langle x_j, x_k\rangle\right| > \epsilon d\right\}\right| \xrightarrow{\nu\to\infty} 0. \tag{17}$$

Here, n, d, and the x_i depend on a hidden index ν such that as ν tends to infinity, so do n and d. A reasonable quantitative analog would be to require A and B above to be bounded, independent of n and d. One could also allow them to grow slowly, as is clear from the statements of the theorems below. Recall that $B \geq \frac{\sigma^2 d}{n}$, so if B is to remain bounded as d tends to infinity, n must tend to infinity at least as fast as d.

In recent work of the author [16], a quantitative version of the Diaconis-Freedman result was proved, giving an explicit bound on $\mathbb{P}\left[d_{BL}(\mu^\theta_{n,d,1}, \gamma_{\sigma^2}) \geq \epsilon\right]$, where γ_{σ^2} is the Gaussian distribution on $\mathbb{R}$ with mean zero and variance σ^2. The results of Section 2 apply immediately to the random vector X uniformly distributed on the n points $\{x_i\}_{i=1}^n$ to give the following k-dimensional extensions.

Theorem 7. *If θ is a random point of $\mathfrak{W}_{d,k}$ and X_θ is distributed according $\mu^\theta_{n,d,k}$, then*

$$d_W(X_\theta, \sigma Z) \leq \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.$$

Theorem 8. *For $\theta \in \mathfrak{W}_{d,k}$, let*

$$d_{BL}(X_\theta, \sigma Z) = \sup_{\|f\|_1\leq 1}\left|\frac{1}{n}\sum_{i=1}^n f\left(\langle x_i, \theta_1\rangle, \ldots, \langle x_i, \theta_k\rangle\right) - \mathbb{E}f(\sigma Z_1, \ldots, \sigma Z_k)\right|;$$

that is, $d_{BL}(X_\theta, \sigma Z)$ is the conditional bounded-Lipschitz distance from X_θ to σZ, conditioned on θ. Then for $\epsilon > 2\pi\sqrt{\frac{B}{d}}$, and θ a random point of $\mathfrak{W}_{d,k}$

$$\mathbb{P}\left[\left|d_{BL}(X_\theta, \sigma Z) - \mathbb{E}d_{BL}(W(\theta), \sigma Z)\right| > \epsilon\right] \le \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{32B}}.$$

Theorem 9. *There is a constant $C > 1$ such that*

$$\mathbb{E}d_{BL}(X_\theta, \sigma Z) \le \frac{C^k B}{d^{\frac{2}{9k+4}}} + \frac{\sigma\sqrt{k}(A+1) + \sigma k}{d-1}.$$

Observe that together, Theorems 8 and 9 show that for $\epsilon \ge \frac{2C^k B}{d^{\frac{2}{5k+4}}} + \frac{\sigma\sqrt{k}(A+1)+\sigma k}{d-1}$,

$$\mathbb{P}\left[d_{BL}(X_\theta, \sigma Z) > \epsilon\right] \le \sqrt{\frac{\pi}{2}} e^{-\frac{d\epsilon^2}{2^7 B}}.$$

Note that the bound on the right tends to zero as $d \to \infty$ for any ϵ in this range. In particular, if A and B are bounded and $\epsilon > 0$ fixed, if $k = c\log(d)$, where c is a sufficiently small constant (depending on ϵ), then $\mathbb{P}\left[d_{BL}(X_\theta, \sigma Z) > \epsilon\right]$ decays exponentially as d tends to infinity.

4. Appendix: The covering number of the class $C^m_M(\mathfrak{X})$

Consider the class $C^m(\mathfrak{X})$ of m-times continuously differentiable functions on $\mathfrak{X} \subseteq \mathbb{R}^d$ with norm defined by

$$\|f\|_m = \sup_{0\le k\le m} \sup_{x\in\mathfrak{X}} \|D^k f(x)\|_{op}.$$

Here, $D^k f(x)$ denotes the symmetric k-linear form given in components by

$$D^k f(x)(y_1, \ldots, y_k) := \sum_{i_1,\ldots,i_k} \frac{\partial^k f}{\partial x_{i_1}\cdots\partial x_{i_k}}(x) y_1^{i_1}\cdots y_k^{i_k},$$

where $y_j = (y_j^1, \ldots, y_j^d)$. For an intrinsic definition of $D^k f(x)$, see Federer [9].

Let $C^m_M(\mathfrak{X})$ be the ball of radius M of $C^m(\mathfrak{X})$ with respect to $\|\cdot\|_m$; in this section, the ϵ-covering number of $C^m_1(\mathfrak{X})$ with respect to the norms $\|\cdot\|_\infty$ (defined the usual way; in our notation, this is $\|\cdot\|_0$) and $\|\cdot\|_1$ is calculated for $m \ge 2$. The proof closely follows the approach in van der Vaart and Wellner [22] but uses the definition of $D^k f$ as a k-linear form instead of working in coordinates with the partial derivatives of f.

First, choose a δ-net $\{y_i\}_{i=1}^n$ of $\mathfrak{X}$, with $\delta = \delta(\epsilon)$ to be determined. One can choose such a net so that $n \le \frac{\mathrm{vol}(\mathfrak{X}_1)}{\delta^d}$, where $\mathfrak{X}_1 := \{x \in \mathbb{R}^d : \inf_{y\in\mathfrak{X}} |x-y| \le 1\}$. Now, associate to each $f \in C^m_1(\mathfrak{X})$ an $(m-1)\times n$ array of operators in the following way. In the space of symmetric k-linear forms on $\mathbb{R}^d$, choose a $\frac{\delta^{m-k}}{2}$-net $\{T_i\}_{i=1}^M$, with respect to the operator norm. The (i,j)-th entry of the array A_f associated to f is chosen to be the closest point in the appropriate net to the i-linear form $D^i f(y_j)$. One can choose δ_0 and δ_1 such that if $f, g \in C^m_1(\mathfrak{X})$ have $A_f = A_g$ (with respect to either the δ_0 or the δ_1 nets), then $\|f-g\|_\infty \le \epsilon$ for δ_0 and $\|f-g\|_1 \le \epsilon$ for δ_1, as follows. For $x \in \mathfrak{X}$ given, choose y_i with $|x - y_i| \le \delta$. By Taylor's theorem applied to $f-g$,

$$(f-g)(x) = \sum_{k=0}^{m-1} \frac{1}{k!}\left\langle D^k(f-g)(y_i), (x-y_i, \ldots, x-y_i)\right\rangle + R,$$

with $|R| \le \frac{2|x-y_i|^m}{m!} \le \frac{2\delta^m}{m!}$. Since $A_f = A_g$, it follows that $\|D^k(f-g)(y_i)\|_{op} \le \delta^{m-k}$ for $1 \le k \le m-1$, thus by the expansion above,

$$\begin{aligned}|(f-g)(x)| &\le \sum_{k=0}^{m-1} \frac{\delta^k}{k!}\|D^k(f-g)(y_i)\|_{op} + \frac{2\delta^m}{m!}\\ &\le \sum_{k=0}^{m-1} \frac{\delta^k}{k!}\delta^{m-k} + \frac{2\delta^m}{m!}\\ &< 5\delta^m,\end{aligned}$$

since $m \ge 1$. It follows that choosing $\delta_0 = \left(\frac{\epsilon}{5}\right)^{\frac{1}{m}}$ means that if $A_f = A_g$ then $\|f-g\|_\infty < \epsilon$.

To choose δ_1, apply Taylor's theorem to $D(f-g)$: if $|v| = 1$,

$$\langle D(f-g)(x), v\rangle = \sum_{k=1}^{m-1} \frac{1}{k!}\left\langle D^k(f-g)(y_i), (x-y_i, \dots, x-y_i, v)\right\rangle + R,$$

(with $x - y_i$ occurring $k-1$ times), and $|R| \le \frac{2|x-y_i|^{m-1}|v|}{m!} \le \frac{2\delta^{m-1}}{m!}$. As above, this implies

$$|D(f-g)(x)| \le \sum_{k=1}^{m-1} \frac{\delta^{k-1}}{k!}\|D^k(f-g)(y_i)\|_{op} + \delta^{m-1} < 5\delta^{m-1},$$

and thus $\|f-g\|_1 \le \epsilon$ if $\delta_1 = \left(\frac{\epsilon}{5}\right)^{\frac{1}{m-1}}$.

To bound the size of an ϵ-net for $C_1^m(\mathcal{X})$, it now only remains to count the number of possible arrays A_f for $f \in C_1^m(\mathcal{X})$. Begin by counting the number of possibilities for the first column. Since $D^k f(y_1)$ is approximated in the k-1 entry of A_f by a point from a $\frac{\delta^{m-k}}{2}$-net, the size of such a net is needed. The space of symmetric k-linear forms is a finite-dimensional normed space, and the size of a net for the unit ball of such a space is given in Milman and Schechtman [17], in terms of the dimension of the space. To define an element T of this space, it suffices to define $T(e_1, \dots, e_1, \dots, e_d, \dots, e_d)$, where e_j appears k_j times with $k_j \ge 0$ for each j and $\sum_{j=1}^d k_j = k$. The number of such vectors (k_j) is well-known (see, e.g., [20]) to be $\binom{k+d-1}{k}$. It follows from the bound in [17] that there is a $\frac{\delta^{m-k}}{2}$-net of the space of symmetric k-linear forms of size not greater than $\left(1 + \frac{4}{\delta^{m-k}}\right)^{\binom{k+d-1}{k}} \le \left(\frac{5}{\delta^{m-k}}\right)^{\binom{k+d-1}{k}}$, assuming that $\delta < 1$. Since the only interesting case is $\epsilon \le 2$ (since $\|f-g\|_i \le 2$ for $i = 0, 1$ automatically), this is no restriction. The number of possibilities for the first column of A_f

is thus bounded by

$$\begin{aligned}
\prod_{k=0}^{m-1}\left(\frac{5}{\delta^{m-k}}\right)^{\binom{k+d-1}{k}} &\leq \prod_{k=0}^{m-1}\left(\frac{5}{\delta^{m-k}}\right)^{\frac{(m+d-2)^k}{k!}} \\
&= \left(\frac{5}{\delta^{m}}\right)^{\sum_{k=0}^{m-1}\frac{(m+d-2)^k}{k!}}\left(\delta\right)^{\sum_{k=0}^{m-1}k\frac{(m+d-2)^k}{k!}} \\
&\leq \left(\frac{5}{\delta^{m}}\right)^{e^{m+d-2}}\left(\delta\right)^{m+d-2} \\
&= \exp\left[\left(\log(5)-m\log(\delta)\right)e^{m+d-2}+(m+d-2)\log(\delta)\right] \\
&\leq \exp\left[\left(\log(5)-m\log(\delta)\right)e^{m+d-2}\right],
\end{aligned}$$

since $\delta<1$ and $m+d-2\geq 0$. To bound the number of possibilities in the remaining columns, assume that the y_i have been ordered such that for all $j>1$, there is an $i<j$ with $|y_i-y_j|<2\delta$. Now, for unit vectors $v_1,\ldots,v_k\subseteq\mathbb{R}^d$, define the function $F(x):=\left\langle D^kf(x),(v_1,\ldots,v_k)\right\rangle$, where the dependence of F on the v_i has been suppressed. By Taylor's theorem,

$$F(y_j)=\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\left\langle D^\ell F(y_i),(y_j-y_i,\ldots,y_j-y_i)\right\rangle+R,$$

with $|R|\leq\frac{(2\delta)^{m-k}}{(m-k)!}$. Let $A_f(i,j)$ denote the i-j-th entry of the array A_f. Then

$$\begin{aligned}
&\left|F(y_j)-\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\langle A_f(i,k+\ell),(y_j-y_i,\ldots,y_j-y_i,v_1,\ldots,v_k)\rangle\right| \\
&\quad\leq(2\delta)^{m-k}+\left|\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\left\langle D^\ell F(y_i),(y_j-y_i,\ldots,y_j-y_i)\right\rangle\right. \\
&\qquad\qquad\left.-\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\langle A_f(i,k+\ell),(y_j-y_i,\ldots,y_j-y_i,v_1,\ldots,v_k)\rangle\right| \\
&\quad=(2\delta)^{m-k}+\left|\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\left\langle D^{\ell+k}f(y_i),(y_j-y_i,\ldots,y_j-y_i,v_1,\ldots,v_k)\right\rangle\right. \\
&\qquad\qquad\left.-\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\langle A_f(i,k+\ell),(y_j-y_i,\ldots,y_j-y_i,v_1,\ldots,v_k)\rangle\right| \\
&\quad\leq(2\delta)^{m-k}+\sum_{\ell=0}^{m-1-k}\frac{1}{\ell!}\left\|D^{k+\ell}f(y_i)-A_f(i,k+\ell)\right\|_{op}(2\delta)^\ell \\
&\quad\leq(2\delta)^{m-k}\left(1+\frac{e}{2}\right).
\end{aligned}$$

That is, given the information in the previous columns, the symmetric k-linear form $T(v_1,\ldots,v_k):=\left\langle D^kf(y_i),v_1,\ldots,v_k)\right\rangle$ is within a ball of radius $(2\delta)^{m-k}\left(1+\frac{e}{2}\right)$ with respect to the operator norm. By the same argument that bounds the size of the original $\frac{\delta^{m-k}}{2}$-net in the space, the

number of points of the net within this ball of radius $(2\delta)^{m-k}\left(1+\frac{\epsilon}{2}\right)$ is bounded by

$$\left(1+\frac{4(2\delta)^{m-k}\left(1+\frac{\epsilon}{2}\right)}{\delta^{m-k}}\right)^{\binom{k+d-1}{k}}=\left(1+2^{m-k+2}\left(1+\frac{\epsilon}{2}\right)\right)^{\binom{k+d-1}{k}}\le\left(2^{m-k+4}\right)^{\binom{k+d-1}{k}}.$$

It follows that the number of possibilities for the column entries of A_f after the first column is specified is bounded by

$$\begin{aligned}
&\left[\prod_{k=0}^{m-1}\left(2^{m-k+4}\right)^{\binom{k+d-1}{k}}\right]^{\frac{\mathrm{vol}(\mathcal{X}_1)}{\delta^d}}\\
&\qquad=\exp\left[\frac{\mathrm{vol}(\mathcal{X}_1)}{\delta^d}\sum_{k=0}^{m-1}\binom{k+d-1}{k}\big((m-k+4)\log(2)\big)\right]\\
&\qquad\le\exp\left[\frac{\mathrm{vol}(\mathcal{X}_1)\big((m+4)\log(2)\big)}{\delta^d}\sum_{k=0}^{m-1}\frac{1}{k!}(d+m-2)^k\right]\\
&\qquad\le\exp\left[\frac{\mathrm{vol}(\mathcal{X}_1)\big((m+4)\log(2)\big)}{\delta^d}e^{d+m-2}\right].
\end{aligned}$$

It now follows that the total number of possible entries of A_f is bounded by

$$\exp\left[\left(\log(5)-m\log(\delta)+\frac{\mathrm{vol}(\mathcal{X}_1)\big((m+4)\log(2)\big)}{\delta^d}\right)e^{d+m-2}\right].$$

Recall that δ_0 and δ_1 were chosen such that $\delta_0=\left(\frac{\epsilon}{5}\right)^{\frac{1}{m}}$ and $\delta_1=\left(\frac{\epsilon}{5}\right)^{\frac{1}{m-1}}$. The ϵ-covering number (for $\epsilon<2$) of $C_1^m(\mathcal{X})$ with respect to $\|\cdot\|_\infty$ is thus bounded by

$$\exp\left[\left(2\log(5)-\log(\epsilon)+\frac{c_0}{\epsilon^{\frac{d}{m}}}\right)e^{d+m-2}\right]$$

with

$$c_0=\mathrm{vol}(\mathcal{X}_1)\big((m+4)\log(2)\big)(5)^{\frac{d}{m}}.$$

The ϵ-covering number of $C_1^m(\mathcal{X})$ for $\epsilon<2$ and $m\ge2$ with respect to $\|\cdot\|_1$ is bounded by

$$\exp\left[\left((3\log(5)-\frac{m}{m-1}\log(\epsilon)+\frac{c_1}{\epsilon^{\frac{d}{m-1}}}\right)e^{d+m-2}\right]$$

with

$$c_1=\mathrm{vol}(\mathcal{X}_1)\big((m+4)\log(2)\big)(5)^{\frac{d}{m-1}}.$$

Acknowledgements: Many thanks to Michel Ledoux, Mark Meckes, and Luke Rogers for helpful comments and suggestions.

References

[1] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989.

[2] Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen-Stein method. *Statist. Sci.*, 5(4):403–434, 1990. With comments and a rejoinder by the authors.

[3] S. G. Bobkov. On concentration of distributions of random weighted sums. *Ann. Probab.*, 31(1):195–215, 2003.

[4] Sourav Chatterjee and Elizabeth Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008.

[5] Louis H. Y. Chen. Poisson approximation for dependent trials. *Ann. Probability*, 3(3):534–545, 1975.

[6] Persi Diaconis. Stein's method for Markov chains: first examples. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 27–43. Inst. Math. Statist., Beachwood, OH, 2004.

[7] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 1984.

[8] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[9] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[10] F. Götze and A. N. Tikhomirov. Limit theorems for spectra of random matrices with martingale structure. In *Stein's method and applications*, volume 5 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 181–193. Singapore Univ. Press, Singapore, 2005.

[11] Susan Holmes. Stein's method for birth and death chains. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 45–67. Inst. Math. Statist., Beachwood, OH, 2004.

[12] Lars Hörmander. *The analysis of linear partial differential operators. I*. Classics in Mathematics. Springer-Verlag, Berlin, 2003. Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)].

[13] Wei-Liem Loh. Stein's method and multinomial approximation. *Ann. Appl. Probab.*, 2(3):536–554, 1992.

[14] H.M. Luk. Stein's method for the gamma distribution and related statistical applications. 1994. Doctoral dissertation, University of Southern California.

[15] Elizabeth Meckes. On Stein's method for multivariate normal approximation. 2009. To appear in *High Dimensional Probability V*.

[16] Elizabeth S. Meckes. Quantitative asymptotics of graphical projection pursuit. *Electron. Commun. Probab.*, 14:176–185, 2009.

[17] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[18] A. Pickett. Rates of convergence of χ^2 approximations via Stein's method. 2004. Doctoral dissertation, University of Oxford.

[19] G. Reinert and A. Röllin. Multivariate normal approximation with stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.*, 37(6):2150–2173, 2009.

[20] Sheldon Ross. *A first course in probability*. Macmillan Publishing Co., Inc., New York, 1976.

[21] V. N. Sudakov. Typical distributions of linear functionals in finite-dimensional spaces of high dimension. *Dokl. Akad. Nauk SSSR*, 243(6):1402–1405, 1978.

[22] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[23] Heinrich von Weizsäcker. Sudakov's typical marginals, random linear functionals and a conditional central limit theorem. *Probab. Theory Related Fields*, 107(3):313–324, 1997.